OVERVIEW OF THE HCUP STATE INPATIENT DATABASES

TABLE OF CONTENTS

OVERVIEW

The release of the Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID) is made possible through a Federal-State-Industry partnership to build a multi-State health care data system. This partnership is sponsored by the Agency for Healthcare Research and Quality (AHRQ), formerly known as the Agency of Health Care Policy and Research (AHCPR). The HCUP databases are managed by the Center for Organization and Delivery Studies.

In general, the SID contain the universe of that state's hospital inpatient discharge records. They are composed of annual, state-specific files that share a common structure and common data elements. Most data elements are coded in a uniform format across all states. In addition to the core set of uniform data elements, the SID include state-specific data elements or data elements available only for a limited number of states. The uniform format of the SID helps facilitate cross-state comparisons. In addition, the SID are well suited for research that requires complete enumeration of hospitals and discharges within market areas or states.

HOW THE SID DIFFER FROM STATE DATA FILES

The SID available through the HCUP Central Distributor differ from the data files available from the Data Organizations in the following ways:

- data elements available on the files, and
- coding of data elements.

Because the Data Organizations dictate which data elements may be released though the HCUP Central Distributor, the data elements on the SID are a subset of the data collected by the corresponding Data Organizations. HCUP uniform coding is used on most data elements on the SID. A few state-specific data elements retain the original values provided by the respective Data Organizations.

TYPES OF HOSPITALS INCLUDED

The hospitals included in the SID depend on the information provided by the Data Organizations. Most state government data organizations provide information on all acute care hospitals in the respective state. Private data organizations are often restricted to member hospitals and may not provide information on all hospitals in their state.

IDENTIFYING HOSPITALS

Up to three hospital identifiers are on the SID.

- The HCUP-specific hospital identifier is always included on the SID, but is coded only for community hospitals as defined by the American Hospital Association (AHA) Annual Survey of Hospitals. The AHA Annual Survey definition of a community hospital includes nonfederal short-term hospitals whose facilities are available to the public. Short-term hospitals are defined as hospitals with an average length of stay less than 30 days. Both general and specialty hospitals (e.g., obstetrics and gynecology, rehabilitation,

orthopedics, and eye, ear, nose and throat) are included.

- Some Data Organizations allow the AHA hospital identifier to be included on the SID. This data element enables the SID to be linked to the AHA Annual Surveys of Hospitals which contain information on hospital characteristics. Similar to the HCUP-specific hospital identifier, the AHA hospital identifier is coded only on community hospitals.

- Some Data Organizations allow the original hospital identifier they collect to be included on the SID. If available on the SID, this identifier is coded for all hospitals and may distinguish different units within a hospital.

The Technical Supplement on *Mapping Source-Specific Hospital Identifiers to AHA Hospital Identifiers* describes the mechanism used during HCUP processing to reconcile the differences between the state and AHA hospital identifiers.

SID FILE STRUCTURE

Based on the availability of data elements across states, data elements included in the HCUP State Inpatient Databases are separated into three files:

- Core File,

- State-Specific File, and

- AHA Linkage File.

The AHA Linkage File is not available for all states. The Data Organizations in participating states determine the release of data elements in each type of file.

CORE FILES

Files containing *core data elements* form the nucleus of the HCUP State Inpatient Databases. Core data elements meet at least one of the following criteria:

- they are available from all or nearly all data sources;

- they lend themselves to uniform coding across sources; or

- they are needed for day-to-day applications (e.g., length of stay, patient age).

The Core Files are discharge-level files that are sorted by the state-defined hospital identifier (DSHOSPID) and the unique SID record identifier (SEQ_SID). If the Data Organization has not allowed the release of DSHOSPID, the sort key is SEQ_SID alone.

STATE-SPECIFIC FILES

Files containing *state-specific data elements* are intended for limited use. State-specific data elements meet at least one of the following criteria:

- they are available from a limited number of sources;
- they do not lend themselves to uniform coding across sources; or
- they are not needed for day-to-day applications.

The State-Specific Files are discharge-level files that are sorted by the state-defined hospital identifier (DSHOSPID) and the unique SID record identifier (SEQ_SID). If the Data Organization has not allowed the release of DSHOSPID, the sort key is SEQ_SID alone. The State-Specific Files contain the same discharges included in the Core Files, but have different data elements.

AHA LINKAGE FILES

Files containing *AHA linkage data elements* allow the HCUP SID to be used in conjunction with the American Hospital Association (AHA) Annual Survey of Hospitals data files. These files contain information about hospital characteristics and are available for purchase through the AHA.

The AHA Linkage Files are hospital-level files with one record for each state-defined hospital identifier (DSHOSPID). These files are sorted by DSHOSPID.

Data Organizations determine whether the AHA linkage data elements will be provided with the HCUP SID. Not all Data Organizations allow the linkage data elements.

COMBINING INFORMATION ACROSS SID FILES

Both the Core File and the State-Specific File are discharge-level files. To combine data elements across files, merge the two files by the unique record identifier SEQ_SID. There will be a one-to-one correspondence of records. Both files contain the same discharges, but have different data elements.

Combining the Core or State-Specific Files (which are discharge-level files) with the AHA Linkage File (which is a hospital-level file) needs to be handled carefully. For example, the Core and State-Specific Files may contain 5,000 discharges for DSHOSPID “A”, but the AHA Linkage File contains only 1 record for DSHOSPID “A”. The files can be merged by DSHOSPID as long as the different levels of aggregation are considered.

HCUP CODING

The following objectives guided the definition of data elements included in the HCUP State Inpatient Databases:

- Make the database as usable as possible without extensive editing by analysts.
- Retain the largest amount of information available from the original sources, while still maintaining consistency among sources.
- Structure the information for efficient storage, manipulation, and analysis.
- Set data element attributes (type and length) to accommodate all expected discharge data. The required characteristics were determined from:

- The actual characteristics of state and hospital association data tabulated in the HCUP Feasibility Study (*AHCPR Hospital Cost Database Feasibility Study*, Contract No. 282-90-0029).
- National standards, including the Uniform Hospital Discharge Data Set (UHDDS), Uniform Bill 1982 (UB-82), and Uniform Bill 1992 (UB-92).

Coding of Data Elements

Data elements are coded as shown in the following table:

Coding Conventions	
Values have been:	**Examples of data elements:**
Retained in the form provided by the data source	Diagnosis and procedure codes
Encrypted into synthetic values	Physician identifiers, person identifiers
Recoded into uniform coding schemes	Sex, race, expected primary pay source
Calculated (when possible)	Age, length of stay, day of principal procedure
Assigned using external algorithms	Diagnosis Related Groups (DRGs), Clinical Classifications Software (CCS), formerly known as Clinical Classifications for Health Policy Research (CCHPR)

Attributes of Data Elements

Data elements are defined as numeric or character.

- Numeric format is used for data elements that are reasonable to express numerically (e.g., age of patient); and for most categorical data elements (e.g., sex of patient).

  Categorical data elements are expressed in numeric format, because that format:
  - facilitates logical comparisons of indicator data elements and
  - permits flexibility in the creation of summary statistics.

- Character format is used for data elements that contain alphanumeric characters not amenable to recoding. Some data elements are expressed in character format because:
  - the alphanumeric data have a recognized significance that must be preserved (e.g., ICD-9-CM diagnosis and procedure codes); and
  - there is no reasonable conversion to numeric coding (e.g., encrypted physician identifiers).

- To save storage space, data element lengths are limited to what is necessary to accommodate the expected data.

Missing Values

Special missing values have been used in HCUP data elements to indicate details of data availability and quality. Missing values differ depending on whether you have obtained HCUP data in SAS transport or EBCDIC/ASCII formats.

- ***Missing Data***

  When:
  - the source has defined an explicit value as unknown or unavailable
  - the source uses a default missing value to indicate missing data
  - exploratory statistics show an undocumented value with a frequency suggestive of a missing value, *and* it is a commonly used missing value (e.g., blank, zero, or 9-filled), or when contacted, the source confirms that the value is unknown or unavailable

  The following missing values are assigned:

  SAS
  - a value of "." for numeric data elements
  - " " (blank) for character data elements

  EBCDIC/ASCII
  - a negative 9-filled value (-9, -99, -999, etc.) for numeric data elements
  - " " (blank) for character data elements

- ***Invalid Data***

  When the source data contain undocumented, out-of-range, or invalid values, e.g., a negative value for age, or an alpha character in a numeric field, the following missing values are assigned:

  SAS
  - ".A" for numeric data elements

  EBCDIC/ASCII
  - a negative 8-filled value (-8, -88, etc.) for numeric data elements

  For diagnoses and procedures, an invalid code is retained in the diagnosis/procedure data element. The presence of the invalid code is recorded in the indicator data element (DXVn/PRVn) associated with that diagnosis or procedure.

- ***Data Unavailable from Source***

  When a source does not provide a data element, the following missing values are assigned:

  SAS
  - ".B" for numeric data elements

  EBCDIC/ASCII
  - a negative 7-filled value (-7, -77, etc.) for numeric data elements

To conserve space on the publicly released SID files, data elements that were unavailable from the source, i.e., coded as .B for all records in a year, were excluded.

- ***Inconsistent Data***

  Related data elements within the same record were checked for logical consistency, e.g., a procedure of *hysterectomy* reported with a sex of *male* is inconsistent. When such inconsistencies were identified, the following missing values were assigned:

  SAS
  - ".C" for numeric data elements

  EBCDIC/ASCII
  - a negative 6-filled (-6, -66, etc.) value for numeric data elements

  For diagnoses and procedures, an inconsistent code is retained in the diagnosis/procedure data element. The presence of the inconsistent code is recorded in the indicator data element (DXVn/PRVn) associated with that diagnosis or procedure.

See the *Quality Control* section below for details.

QUALITY CONTROL

This section describes the procedures used to assess data quality for each data source participating in HCUP.

Quality Control Philosophy

Edit procedures were applied to HCUP data. Editing followed explicit rules:

- Make the data usable without extensive further editing.
- Confirm that data values are valid, internally consistent, and consistent with established norms, when feasible.
- Use some edit procedures to set questionable and inconsistent values to inconsistent (.C or negative 6-filled). Use other edit procedures only to tabulate edit failures. Use the latter to evaluate whether systematic problems exist.
- Never "fix" or impute data. Set invalid or inconsistent values to missing or, for diagnoses and procedures, set flags to indicate invalid or inconsistent codes. This preserves the analyst's ability to investigate data anomalies.
- Some data elements are more important than others because:
  - they are coded more reliably because they relate to reimbursement; and
  - without these data elements, a discharge record is not useful for most analytic purposes.

  Therefore, values of these data elements should be retained even in the presence of

conflicting information. In order of importance, these data elements are:

1. Discharge date (and within discharge date: year, month, and day)
2. Admission date
3. Principal diagnosis

- Tabulate instances of edit failures and use these to assess data quality for each data source.

Quality Review

The following statistics were reviewed by an independent contractor for each year and data source (or for each different layout if the source changed file layouts during the year):

- For all numeric data elements — means, number of missing and nonmissing values, minimum, and maximum.
- For all categorical and some continuous data elements — frequency distributions.
- For closely related data elements (e.g., age in years compared to age in days) — cross-frequencies.

For details, see the Technical Supplement on *Quality Control in HCUP Data Processing*.

Automated Quality Control Procedures

The following procedures were applied to each inpatient discharge record:

- *To assess validity of values —*

  For numeric data:
  - Verify numeric data as numeric.
  - Check the range against legal values documented by the data source.
  - Check the range against standard norms (e.g., length of stay is a non-negative value; age in years is between 0 and 124, the maximum allowed by the DRG grouper).
  - Check the values against the maximum allowed for the data element (e.g., length of stay less than 32,767).

  For character data:
  - Verify against norms, when feasible (e.g., diagnosis codes, procedure codes, patient zip codes).

- *To assess internal consistency —*

  Compare values of related data elements (e.g., a procedure of *hysterectomy* should appear with a sex of *female*; admission date should occur *before* discharge date).

  If an inconsistency involves a critical data element (such as discharge date, admission date, or principal diagnosis), retain the critical data element according to the established hierarchy. For example:

- If discharge date falls before admission date, retain discharge date and set admission date and length of stay to inconsistent (negative 6-filled or .C).

- If discharge date is invalid (e.g., February 30), retain discharge quarter and discharge year.

- *To assess consistency with established norms —*

  Compare values to an established norm (e.g., maternal diagnoses should occur with an age between 10 and 55 years).

Data Quality Data Elements

Three types of data quality data elements were created during inpatient discharge data processing.

- **Diagnosis and Procedure Validity Flags — DXVn and PRVn**

  These flags indicate invalid or inconsistent data in the associated diagnosis and procedure data elements. Original values of the diagnoses and procedures are maintained. DXVn and PRVn have the following values:

  0 Diagnosis or procedure code is valid and consistent.

  1 Diagnosis or procedure code is invalid as of the discharge date, plus or minus three months (to allow for anticipation of or lags in response to official ICD-9-CM coding changes).

  .C or -6 Diagnosis or procedure code is inconsistent with age or sex on the same record.

  . or -9 No diagnosis or procedure coded.

- **Neonatal/Maternal Indicator Flag — NEOMAT**

  This data element identifies discharges with neonatal and/or maternal diagnoses or procedures. Maternal diagnoses and procedures do not necessarily result in a delivery. For a definition of neonatal and maternal diagnoses and procedures, refer to the Technical Supplement on *Quality Control in HCUP Data Processing*.

  NEOMAT has the following values:

  0 No neonatal/maternal diagnoses or procedures.

  1 Maternal diagnoses and/or procedures are present.

  2 Neonatal diagnoses are present.

  3 Both neonatal and maternal diagnoses and/or procedures are present.

- **Edit-Check Data Elements — ED010-ED952**

  These binary data elements identify inconsistencies between related data elements on the same record. The data elements have the following values:

  0 The problem was not found, or the edit check was not applicable.

  1 The record failed the edit check.

To conserve space on the publicly released SID files, the edit check data elements were not included. For details on consistency and edit checks performed during the HCUP data processing, refer to the Technical Supplement on *Quality Control in HCUP Data Processing*.